

DRAM CORE REFRESH WITH REDUCED SPIKE CURRENT

This application claims priority to the provisional patent application entitled, "DRAM Core Refresh with Reduced Overhead", Serial Number 60/061,767, filed October 10, 1997.

5

BACKGROUND OF THE INVENTION

The present invention relates to dynamic random access memories (DRAMs), and in particular to refreshing techniques.

DRAM devices function by storing charge on a capacitor at each memory location. The capacitor loses the charge over time, and accordingly it needs to be recharged to its original level, whether a 1 or 0, periodically. This period is known as the refresh period, tREF. A typical tREF value is 64 ms. This recharging is termed refreshing, and is done by doing a row access for every row in the memory device. In newer DRAM designs, the rows and columns are broken into multiple banks to break down large DRAM arrays into smaller pieces.

The row access operation is performed in two steps. First, a row is opened, or turned on, in a "sense" operation, by turning on the sense amps. Next, the row is closed by doing a "precharge" operation, using a precharge equalization transistor. When a refresh is performed, there is no need to select a particular bit with a column address, and the data is not read.

As the storage capacity of memory devices increases and the number of banks per device increases, the issue rate of refresh commands sent to the memory device from the controller also increases and can introduce a refresh overhead that unacceptably impacts the performance of normal memory accesses.

25

To reduce refresh overhead, it is desirable to refresh more than one bank for each refresh command. This approach can be called multibank refresh. With a given refresh command, more than one bank could be either simultaneously or sequentially refreshed. From a performance standpoint, it is more desirable to refresh banks simultaneously, so that the time that bank resources are tied up is minimized.

Simultaneous multibank refresh, however, has the problem of current spikes. Each refresh operation for each bank requires a certain amount of supply current over time. At the onset of the operation, there is an initial spike of current. This spike is large because the row sensing circuits have been designed to access cell data as quickly as possible in order to minimize the latency to the first allowable page access to bits stored in the sense amps. This spike, characterized by rate of change in current, di/dt , can cause noise problems in a DRAM, since current spikes can reduce the internal supply voltage and cause failure in circuits on the same die or on other devices that share the same supply voltage. With multiple banks simultaneously doing a row sense, the current spike effect can be additive, thus causing greater probability of circuit failure.

Typically, a refresh operation is done by periodically addressing every row with a controller. Thus, interspersed between normal memory access operations, refresh commands are sent in the form of a RAS control signal with a row address. In a prior RAMBUS memory system, these commands are sent in packets which are decoded in the memory chips themselves.

In addition to the normal refreshing interspersed with memory accesses, the memory may also be put into a sleep or stand-by mode. In this mode, it is not being accessed, and minimal power drain is desired. This is accomplished by simply refreshing the memory when needed, and otherwise not doing memory accesses. Memory chips typically have on-chip counters for sequencing through all the rows and banks in order to accomplish such a self-refresh. This allows the memory chips to refresh themselves, without requiring the controller to be turned on to provide it with the refresh addresses and commands.

Because self-refresh mode uses a clock or sequencer on the DRAM itself, a synchronization issue arises when the device comes out of self-refresh mode and the controller takes over memory accesses and controlling refresh. The typical way the synchronization is handled is for the microprocessor, upon powering up out of a self-refresh mode, to send a burst of refresh commands covering all the data locations in the memory chip. In this way, it is ensured that the timing of the last refresh of every memory location is known to the microprocessor.

To keep up with increasing microprocessor speeds, there has been pressure to increase the speed at which memory is accessed. One method for accomplishing this is to shape the current pulse provided by the sense amplifier driver in order to

increase access speed. This is discussed, for example, in an article by H. Geib, W. Rabnd. Schmitt-Landsiedel, entitled "Block-Decoded Sense-Amplifier Driver for High-Speed Sensing in DRAM's", IEEE Journal of Solid-State Circuits, Vol. 23, No. 9, September 1992. As memories become larger, however, more locations need to be
5 refreshed at any given time, and refreshing draws more power in a shorter time for the larger memories. Thus, current spikes can cause significant noise problems on the power line during a refresh operation. Accordingly, refresh protocols limit refreshing to one row at a time in order to control refreshing noise.

10

SUMMARY OF THE INVENTION

The present invention makes multibank refresh more practical by varying the current profile for the row sense and/or row precharge currents during a refresh operation, as compared to normal memory access. Unlike normal memory accesses,
15 data is not needed, and a fast access time is not required. This allows the current to be spread using different circuitry for driving the current so as to lessen current spikes. The spread current is still maintained within the timing of a normal refresh or memory access.

By refreshing multiple banks in response to a single command, the present
20 invention provides a method for reducing the communication overhead over the interface bus to the memory devices for refresh operations.

In one embodiment, the generation of row and bank addresses during refresh is split between on-chip and external commands. A row counter is provided on the memory chip, with the row counter being used for refresh operations, both normal and
25 self-refresh. Only the bank address needs to be sent over the memory bus. Since the same row counter is used for self and normal refresh, only a burst of the banks for a particular row need to be issued to regain synchronization, since the same row counter is used in both modes.

Additionally, in the preferred embodiment, the banks occupy the least
30 significant bits of the address, and the rows occupy the most significant bits of the address. Such an addressing scheme allows the internal row counters to be taken advantage of, and allows a transition from self-refresh to normal refresh with only a burst of the banks.

In one embodiment, the row sense or precharge currents have their profile modified by using multiple transistors in parallel, with only one or some transistors being turned on, or some or all of the transistors being sequentially turned on, so that the full current is not applied at once in the refresh mode. In normal mode, all the
5 transistors can be turned on simultaneously for the fastest access. Alternately, the control signal provided to the sense driver or precharge driver could be ramped at a lower rate, or operate at a lower voltage to limit the current spike of the transistor.

For further understanding of the nature and advantages of the invention,
reference should be made to the following description taken in conjunction with the
10 accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram illustrating row sense current profiles.

15 Fig. 2 is a diagram of a memory core illustrating sense amp organization.

Fig. 3 is a block diagram of a portion of the logic of Fig. 2.

Fig. 4 is a diagram showing the sense amp drivers and pre-charge equalization
transistors in the circuit of Fig. 3.

Fig. 5 is a diagram illustrating the sense amp positive and negative voltage
20 waveforms.

Figs. 6A, 6B and 6C illustrate the circuit and waveforms for parallel drive
transistors.

Figs. 7A and 7B illustrate the circuit and waveforms for a drive transistor which
is ramped.

25 Figs. 8A, 8B, 8C and 8D illustrate the circuit and waveforms for precharge
equalization.

Figs. 9A, 9B and 9C illustrate various control packets.

Fig. 10 is a block diagram illustrating the use of an on-chip row counter.

30 Figs. 11A, 11B and 11C are diagrams of fast and slow clock circuits for low
power binning according to self-refresh power.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

As described above, dynamic random access memory (DRAM) devices require periodic operations to retain data in its storage cells. A refresh operation consists of a row sense operation and a row precharge operation. Each memory cell needs to be refreshed within a given time interval known as tREF, the refresh period. A typical tREF value is 64 ms. As a result, each row in each memory bank needs to be refreshed within tREF.

As the number of banks in memory devices increases and as the number of devices in memory systems increases, the issue rate of refresh commands sent from the controller also increases. This increased traffic due to refresh operations can introduce an overhead that unacceptably impacts performance, as measured by effective data bandwidth and memory access latency.

The present invention implements refreshing of more than one bank for each refresh command to reduce refresh overhead on the memory system bus. This approach is called multibank refreshing. With a given refresh command, either more than one bank is simultaneously refreshed, or more than one bank are sequentially refreshed from the same command. From the performance standpoint, it is more desirable to refresh banks simultaneously, so that the time the bank resources are tied up is minimized.

One of the problems caused by multibank refresh is that of current spikes. Each refresh operation or each bank requires a certain amount of supply current over time. Fig. 1 shows a typical current profile 10 over time for a row sense operation. At a point 12, there is a large initial spike of current near time 0. This spike is large because the row sensing circuits have been designed to access cell data as quickly as possible in order to minimize the latency to the first allowable page access to bits stored in the sense amps. This spike, characterized by rate of change of current, dI/dt , can cause noise problems in a DRAM, since current spikes can reduce the internal supply voltage and cause failure in circuits on the same die or on other devices that share the same voltage supply. With multiple banks simultaneously doing a row sense, the current spike effect can be additive, thus causing greater probability of circuit failure.

The present invention provides methods and apparatus for reducing the current spike to give a profile such as that indicated by profile 14 in Fig. 1. This profile

achieves less of a current spike, but still delivers the required current in approximately the same time. The smaller initial current spike will increase the latency for accessing data, but since data is not accessed in the refresh, this is not a concern for a refresh operation.

5 To understand one embodiment of the invention, a brief overall description of the structure of the memory is useful. Fig. 2 shows the organization of multiple banks in a memory having a number of cell arrays 16 with sense amplifiers 18, and sense amplifier drivers 20.

Fig. 3 illustrates additional logic for one cell array 16. In addition to the sense amplifier driver or logic 20 and sense amplifiers 18, there is additionally shown the bit line precharge circuits 22, with corresponding driver and logic circuit 24. Also shown is a row decoder 26 for selecting particular rows or word lines in cell array 16.

Fig. 4 illustrates some of the details of the sense amplifier driver and bit line precharge circuit 22 (also sometimes called the row precharge circuit). As shown, amplifier driver 20 includes a positive driving PMOS transistor 28, and a negative driving NMOS transistor 30. Transistor 28 is connected to a positive voltage supply 32, Vdd. Transistor 30 is connected to circuit ground 34. When activated, transistor 28 will provide a positive sense amp voltage, SAP, on drive line 36 to the sense amplifiers 18. Drive transistor 30 will provide the negative voltage level, SAN, on line 38 to the sense amplifiers.

Fig. 5 illustrates the voltage levels supplied to the sense amp, identified as SAP voltage waveform 40, and SAN voltage waveform 42. These voltages are applied when a page is opened, or when the sense portion of a refresh operation is initiated.

Bit line precharge circuit 22 includes a precharge equalization transistor 44. After the sense step, this transistor is activated to equalize the lines to the sense amplifiers to close a page. This brings the voltage of both bit lines, as well as the sense amp supply lines 36, 38, to a level of Vdd/2, illustrated as waveform 46 in Fig. 5.

The transitioning of the voltage levels SAP and SAN to their high and low levels as shown in Fig. 5 is what causes the current spike illustrated in Fig. 1. A similar current spike can be caused by the precharge equalization transistor.

Fig. 6A is a circuit diagram of one embodiment of the invention for modifying sense amp driver transistor 30 to reduce the voltage spike. The sense amp driver

transistor 30 (Fig. 4) is replaced by three driver transistors 48, 50 and 52 in parallel. By sequentially turning on these transistors, the current spike can be spread out as illustrated in Fig. 6B. Fig. 6B shows a first waveform SAN0 which will turn on transistor 48 at a time 54. A second signal, SAN1 to the gate of transistor 50, will turn 5 transistor 50 on at a time 56. Finally, a third control signal SAN2 to the gate of transistor 52 will turn transistor 52 on at a time 58. This results in the waveform 60 of Fig. 6B, which has three smaller peaks 62, 64 and 66, instead of the combined current peak of a single transistor. The three transistors can either be of the same size, or different sizes. The first transistor turned on could be the largest transistor, or 10 the smallest. The staggered turn-on would be used for refresh operations only, such as simultaneous multibank refresh, with all three transistors being simultaneously turned on for a normal access operation. Sense logic block 20 of Fig.3 includes logic for determining whether the access is a refresh or normal access, and appropriately staggering the turn-on times or activating all three at the same time. Alternately, only 15 some of the transistors could be used for refresh. Another possibility is to use extra transistors for staggered turn on for refresh, but only a single one of those transistors for a normal access.

A similar circuit is shown in Fig. 6C to replace the positive sense amp circuit 28 with three parallel transistors 68, 70 and 72. Alternately, a different number of 20 transistors than three could be used, or the number of transistors used for the SAP drive line could differ from those used for the SAN drive line.

Figs. 7A and 7B illustrate an alternate embodiment of the invention. Fig. 7A illustrates drive transistor 30 with signal SAN0 on its gate. Fig. 7B illustrates the effect of a lower slew rate and lower voltage on the control line 74 to the gate of transistor 25 30. Line 76 indicates a normal voltage and slew rate for turning on the gate, with its corresponding current profile 78 for the sense amp current, giving the normal spike. Line 80 illustrates a lower slew rate, indicated by ramp portion 82, and the signal also stops at a lower voltage level 84. The resulting profile is illustrated as current profile 86. As can be seen, this is effective in shaping the pulse to spread it. Again, as in 30 the first embodiment, during normal operation voltage and slew rate 76 would be used, while for refresh operations waveform 80 would be used. A similar waveform can be used for the positive sense drive transistor 28, as well as for the precharge equalization transistor 44 of Fig. 4. In other embodiments a reduce current spike can

be achieved using a slower slew rate with the normal voltage level, or by using a lower voltage level (84) with a normal slew rate.

Figs. 8A-8D illustrate similar circuits and waveforms for the precharge transistors. Fig. 8A illustrates standard precharge transistor 44 and sense amplifier

5 18. Fig. 8B illustrates using two precharge transistors 88 and 90, in parallel. As for the sense amp drivers, other numbers of precharged transistors could be used, and they could be of the same size or differing sizes. The gates of the transistors are shown as receiving control signals PCH0 and PCH1. Fig. 8C illustrates staggering the turn-on times of PCH0 and PCH1 to shape the pulse. Fig. 8D illustrates the use
10 of a single transistor as in Fig. 8A, but with a lower slew rate and voltage control signal 92 being used for refresh, as contrasted with a signal 94 used during normal operations.

In one embodiment, the two methods could be combined, with multiple, parallel transistors with staggered turn-ons, and additionally having a lower voltage and slew
15 rate applied to each of the parallel transistors.

Refresh Address Sequencing

As illustrated in Fig. 9A, the present invention reverses the normal address sequencing of the memory core during refresh operations. Both normal refresh and
20 self-refresh have the bank address bits as the LSBs, and the row address as the MSBs. Thus, for each row, all the banks are sequenced through before the next row is addressed. As will be discussed below, an on-chip row counter is used, so that the controller does not have to track row addresses. The sequencing of the bank addresses as the LSB works with the internal counter. The controller does not have
25 to track row addresses during refresh, and only needs to send the bank address during each refresh operation.

This is illustrated in Fig. 9B, which shows a primary control packet for a refresh operation as including a device field 96, a bank field 98, and an op code (refresh)
100. Fig. 9C illustrates the refresh op codes. As shown, there is both a refresh sense
30 command 102, and a refresh precharge command 104 (also called precharge after refresh). These commands are different from a normal sense and precharge command, unlike the prior art which typically would use the same sense and precharge operations for a refresh as for a normal memory access operation.

By utilizing an internal row counter for refresh, the row address need not be transmitted in the control packet. Thus, instead of sending the first packet with the op code, and the second packet with the address, the op code can be inserted into the portion of the packet that would normally have a row address, as illustrated in Fig. 9B, saving control packet bandwidth.

The use of separate refresh commands (i.e., distinct from normal memory access commands) allows the triggering of the separate pulse shaping techniques described above, and also the triggering of the internal row counter. Fig. 10 is a block diagram of the control logic for implementing the refresh operations of the invention.

- 5
- 10 As shown, a control packet decode block 106 will receive the control packet. If the operation is a refresh operation, a control line 108 is asserted. This causes a multiplexer 110 to select an output from the internal row counter 112. In contrast, for normal non-refresh operations, the row address will be provided on line 114 through multiplexer 110 to memory core 116. Bank decode logic 118 will decode the bank
- 15 and provide it on control lines 120 to memory core 116. For a refresh operation, a bank mask 122 or other logic may be used to enable multibank refresh for multiple banks to be simultaneously enabled.

In multiple bank refresh, the appropriate number of MSB's from the bank address can be ignored. For example, a device with 16 banks with 1/8 activation (two 20 banks simultaneously refreshed) would ignore one MSB of the bank address so that a bank address X000 would activate banks 0 and 8. With 1/4 activation, two MSB bits are ignored, and a bank address of XX00 would activate banks 0, 4, 8 and 12. The DRAM design will assign these addresses to banks with dependent bank cores, so that no two adjacent banks are simultaneously activated. Although this multibank 25 addressing scheme naturally supports the most straight-forward sequential address assignment to banks, other assignments are possible.

For the address sequencing, each refresh sense command and each precharge after refresh command includes a bank address. Bank decode logic 118 will compare the bank address to the maximum bank address. When the maximum 30 bank address has been reached, indicating that all banks for a particular row have been refreshed, the bank decode logic 118 sends an increment signal on a line 124 to row counter 112 to increment the row count for the next sequencing of banks. After the maximum row has been sequenced through, the row counter is reset to zero. Thus, by using the bank address as the LSB portion of the address field, all banks are

refreshed before the refresh row address counter is incremented. This arrangement of the row and bank addresses allows simple support for multiple page sizes, since the appropriate number of most significant bits (MSB's) can be ignored.

For self-refresh, the control logic of Fig. 10 includes a bank address counter 125, which generates a sequence of bank addresses at a rate determined by a self-refresh clock 126. When the bank address counter 125 reaches a maximum bank address, indicating that all banks for a particular row have been refreshed, it sends an increment signal on line 124 to row counter 112 to increment the row count for the next sequencing of banks.

5 10 15 20 25 30

Each pulse output by the self-refresh clock 126, in addition to incrementing the bank counter 125, also triggers the refresh driver 128 for driving the sense amplifiers in the memory core so as to perform a refresh operation. The refresh driver 128 is also triggered by the control packet decoder logic 106 when a refresh command is received from the memory controller.

The control packet decode logic 106 outputs a Self-Refresh Enable signal when a self-refresh command is received from the memory controller. The self-refresh enable signal enables operation of the self-refresh clock 126, and is also used as a select signal for a bank address multiplexer 127. Multiplexer 127 sends to the memory core 116 the bank address generated by the bank address counter 125 during self-refresh, and otherwise sends the bank address generated by the bank address decode logic 118.

Self-Refresh Resynchronization

The present invention avoids problems in the prior art of resynchronizing when entering or exiting self-refresh. By using a row counter for both self-refresh and normal refresh, no resynchronization of the rows is needed at all. To handle the resynchronization of the banks, a burst of refresh commands addressing all the banks can be sent by the controller upon the exit from the self-refresh mode. This insures that every bank in the last row has indeed been refreshed before the external controller takes over the task of generating the bank addresses.

Binning

In a low power mode, it is desirable to do self-refresh at a slower frequency, so as to use less power. At the same time, the refresh must be done sufficiently fast enough to maintain the data stored in the DRAM cells. Typically, a clock frequency is chosen which is sufficiently fast so that it will be able to provide the self-refresh timing regardless of variations in device parameters and cell hold time in a particular DRAM. However, some DRAMs may end up having longer cell hold times, and need less frequent refreshes. Accordingly, to take advantage of this, in one embodiment the self-refresh clock 126 uses two internal row counter oscillator circuits, as shown in Fig. 11A. A fast oscillator 130 and a slow oscillator 132 are both provided. Note that fast and slow are relative terms, and the varying speeds can be in fact somewhat close. A multiplexer 134 selects between the two oscillators. After the device is manufactured, the worst cell hold time for the memory cells in a particular DRAM chip is measured, and the fast or slow oscillator is selected by a control signal on line 136, which can be set by blowing a fuse or otherwise to pull it high or low. The control signal on line 136 is a device configuration signal that, in effect, configures the memory device for a particular mode of operation. In this manner, if a particular DRAM chip has a long hold time in its memory cells, the slower clock can be activated for the self-refresh, and the DRAM can be placed into a bin for use in laptop computers or other computers requiring lower power devices.

Figs. 11B and 11C illustrate alternate embodiments. In Fig. 11B, the self-refresh clock 126 uses a single oscillator 138, with a control input 140 (i.e., the aforementioned device configuration signal) varying the speed of oscillation of the single oscillator. In this embodiment, the oscillator 138 may be capable of two or more distinct oscillation speeds, and thus the control (device configuration) signal may be set to any of a corresponding set of two or more predefined signal values. Alternately, in this embodiment the device configuration signal may be implemented as a set of one or more fuses that form part of the self-refresh clock and which determine the frequency of the oscillator, depending on which of the fuses form open circuits and which form closed circuits.

Fig. 11C shows a self-refresh clock having a single fast oscillator 142, with a divider 144 providing the slower clock, which can then be selected through a multiplexer 146.

As will be understood by those of skill in the art, the present invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. Accordingly the foregoing description is intended to be illustrative, but not limiting, of the scope of the invention as set forth in the following
5 claims.